EXHIBIT 79

UNITED STATES DISTRICT COURT

FOR THE NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

RICHARD KADREY, et al., *Individual and Representative Plaintiffs*, v. META PLATFORMS, INC., *Defendant.*	Case No. 3:23-cv-03417-VC

REBUTTAL REPORT OF

CRISTINA VIDEIRA LOPES, PhD

FEBRUARY 03, 2025

65. The employed methodology is almost identical to that of Experiment 3, and shares the same problems, so I will comment on both next. The only difference is that Experiment 2 tested all the models, not just a representative of each model-book. However, that seems to have been pointless, because each of the 5 models for each book produces exactly the same output.[49]

D. REVIEW OF EXPERIMENT 3: NO BASIS FOR CLAIM THAT MODELS HAVE NOT MEMORIZED PLAINTIFFS' WORKS

66. It is a well-documented phenomenon that LLMs memorize training data [10][11][26]. Studies have shown that the level of memorization depends on the number of repetitions and the size of the models [11].

67. Meta, itself, has invested considerable resources to studying this topic.[50] Memorization was an active concern for Llama employees, who feared "exacerbating IP risks."[51] I reviewed documents showing that at least as early as October 2022, Meta employees openly contemplated the risk that Llama would memorize copyrighted books on which it had trained and the company weighed that risk against the benefits of using books as training data, since Meta viewed books as particularly high quality data.[52]

68. As stated by Meta software engineer David Esiobu, "Llama [wa]s annoyingly good at quoting books 🙃,"[53] and Llama's memorization problem was delaying the launch of a new

[49] Ungar Supplemental Material, folder continued_pretraining_in_domain_results.

[50] Meta_Kadrey_00000277 (internal Meta presentation titled "Memorization in LLMs & MME [Memorization Measurement Engine]").

[51] Esiobu Dep. Tr. 200:3-213:11, 89:2-90:22; Meta_Kadrey_00093425; Meta_Kadrey_00079879.

[52] Meta_Kadrey_ 00074729 ("If data quality and volumes are so important for the performance of LLMs, the best resources we can think of are definitely books. … "We should be defining these risks and quantifying it with metrics. Memorization is one piece of the puzzle, I'm sure there's more."); Esiobu Dep. Tr. 178:14-179:6 ("I would say people at Meta were concerned about—myself included, were concerned about these things.").

[53] Meta_Kadrey_00054433; Meta_Kadrey_00054435.

model.[54] After Llama accurately quoted several books in response to "ad-hoc querying," a Meta employee noted "the model is particularly good at quoting books."[55]

69. Several Meta employees assessed both the severity and prevalence of copyrighted content "leakage" highly among other issues with Llama.[56] Meta employees considered memorization among its highest priority issues.[57]

70. Prof. Ungar's report attempts to test memorization using *discoverable extraction* [11], which has become a popular method for studying memorization in LLMs (*e.g.*, [14][5][33][51][59]). When using this method, a piece of data is considered memorized by an LLM when it can be split into two parts such that using the first part as prompt leads the LLM to generate the second part.

71. The most well-known studies of the phenomenon, such as those cited above, underestimate the true extent of memorization in LLMs, because they rely on single-sequence greedy sampling [26] —*i.e.,* they use LLMs as deterministic models that always generate the most likely next token. As stated before, that is not how LLMs are configured for use. Hayes et al. [26] showed that taking the non-deterministic nature of LLMs into account can reveal cases of higher memorization rates compared to rates found through greedy extraction. Extraction of training data from LLMs is an active area of research because of its security and privacy implications. If training data is extractable by some procedure, it means that the model, somehow, retains copies of that training data, and those copies can potentially be accessed by anyone with access to the model.

72. While there is consensus that LLMs memorize data, and that such memorization is a security, privacy, and legal risk, there is no known technique currently capable of extracting the

[54] Meta_Kadrey_00080661 ("Memorization (IP) and PII mitigations require new pretraining before public launch…").

[55] Meta_Kadrey_00214257 ("Notebook – Memorization in Genesis").

[56] Meta_Kadrey_00211621 (2023 H2 priorities tracker).

[57] *See*, *e.g.*, Meta_Kadrey_00063146 ("North Star: Ensure that models do not memorize and regurgitate training data."); Meta_Kadrey_00094461 (rating "IP/copyright leakage" as "p0"); Esiobu Dep. Tr. 218:4-5 (defining "p0"); Meta_Kadrey_00049649 ("Currently it is possible that longer contexts which occur fewer times in pre-training are just as likely to get memorized as shorter token sequences that occur more times, especially for large models like the 405B one[.]").

memorized data reliably. However, theoretical models show that with enough attempts, memorized data is extractable and can be revealed [26]. To the extent that the Llama models store the training data in their weights, possibly in the form of statistics that can be explored, it is just a matter of time until someone discovers a procedure for successful extraction.

1. DESCRIPTION OF THE EXPERIMENT

73. Experiment 3 was a discoverable memorization experiment along the lines of all the others cited above. The goal here was to measure the Llama models' discoverable memorization of 14 of the Plaintiffs' works. The procedure was similar to the one described for Experiment 2.

74. For each book, 100 passages were selected.[58] The models were given the first 150 tokens of those passages and prompted to complete the next 50. The prompt included the instruction preamble before the passage. I detected the same pattern on those passages as that in Experiment 2, namely, a certain set of 100 passages of each book were used for Llama 2 experiments, while a different set of 100 passages were used for experiments with the Llama 3 models.[59] Again, no explanation is provided for using two sets of prompts.

75. For each completion from the models, the experimenters counted the number of tokens in the sequence that were the same as the ground truth until the first token that was different.

76. The main report shows averages for Llama 2-7B, Llama 3-8B and Llama 3.1-8B. The supplemental data files also contain model responses and aggregated results for Llama 2-13B, Llama 2-70B, Llama 3.2-1B, Llama 3.2-3B, and Llama 3-70B.

77. For all the Llama 2 models, the data files contain additional responses and results from prompts that conform to the common discoverable extraction studies, *i.e.*, no instruction. Responses and results for this prompting style for Llama 3 models were not included.

78. The results in Prof. Ungar's Opening Report for both Experiment 2 and Experiment 3 show that for all the tested models and for all the tested books, the number of tokens extracted was, on average, very small: less than 2 tokens.

[58] Ungar Opening Report ¶ 191.

[59] Ungar Supplemental Materials, folder plaintiff_continuation_results.

79. However, the detailed data files show continuations much higher than Prof. Ungar included in his report, including continuations as high as 50 tokens, *i.e.*, complete verbatim completion.

2. ASSESSMENT

80. Being that the possible number of passages from each book is in the tens of thousands, 100 samples from each book corresponds to a confidence interval less than 70%±5—too uncertain to draw strong conclusions.[60] Such a low confidence interval is unusual in memorization experiments in contemporary practice. 100 was used in one of the earliest memorization experiments at Google [14], but from then on, the number of samples of these experiments increased considerably. Carlini et al. used 50,000 samples in one experiment and 1,000 samples in each repetition bin in another [11]. Google later used 2,000 [33] and 10,000 samples [5][51][59]. This is a technicality, but, in keeping with best practices, it is indicative that a more robust analysis of the results would be needed.

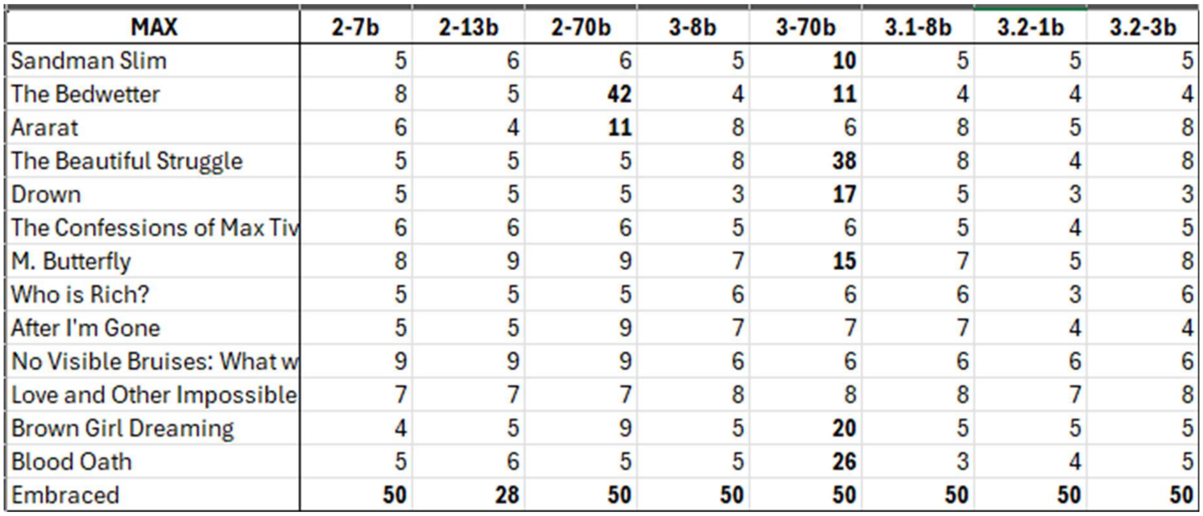
81. To make matters worse, Prof. Ungar reports only on *average* token completions.[61] It is well-known that averages are misleading especially when the data is highly skewed, as is the case here. For example, as demonstrated upon closer examination of the data, the averages conceal examples of significant completions of Plaintiff works which exceed double-digit tokens.

82. **Table 4** shows the longest length of completions for each of the 14 books, with double digits highlighted. This table was not in Prof. Ungar's reports. It was compiled by me by looking at the data files he provided as supplemental material. This table shows the longest completions that each model generated. In this case, longest completions are more meaningful than averages, because we are looking for the existence of memorization of specific works, not for general model behavior with respect to memorization. Thus, even one single occurrence of significant regurgitation of a specific book must be reported as it is proof of existence of memorization. Prof. Ungar failed to report his own results that clearly illustrate meaningful memorization by the Llama models.

[60] See, for example, https://www.calculator.net/sample-size-calculator.html

[61] Ungar Opening Report, pages 217-218, Tables 36 & 37.

83. Consistent with results reported in the literature [11] and in internal Meta experiments,[62] here too, the largest models (70B) show more memorization than the smaller ones. I expect the Llama 405B models to exhibit even stronger signs of memorization. Indeed, Prof. Ungar's experiments are additionally flawed because he focuses almost exclusively on the two smallest Llama models, despite the literature uniformly confirming that memorization increases with size. By opting not to run his test on the 405B model, he avoids the highly likely possibility that his experiment would show even higher levels of memorization.



MAX	2-7b	2-13b	2-70b	3-8b	3-70b	3.1-8b	3.2-1b	3.2-3b
Sandman Slim	5	6	6	5	**10**	5	5	5
The Bedwetter	8	5	**42**	4	**11**	4	4	4
Ararat	6	4	**11**	8	6	8	5	8
The Beautiful Struggle	5	5	5	8	**38**	8	4	8
Drown	5	5	5	3	**17**	5	3	3
The Confessions of Max Tiv	6	6	6	5	6	5	4	5
M. Butterfly	8	9	9	7	**15**	7	5	8
Who is Rich?	5	5	5	6	6	6	3	6
After I'm Gone	5	5	9	7	7	7	4	4
No Visible Bruises: What w	9	9	9	6	6	6	6	6
Love and Other Impossible	7	7	7	8	8	8	7	8
Brown Girl Dreaming	4	5	9	5	**20**	5	5	5
Blood Oath	5	6	5	5	**26**	3	4	5
Embraced	**50**	**28**	**50**	**50**	**50**	**50**	**50**	**50**

Table 4. Longest completions--signs of memorization of Plaintiffs' books in Prof. Ungar's experiment

84. Looking closer, these long completions are neither trivial nor common. **Table 5** shows some examples:

[62] Meta_Kadrey_00000277 (internal Meta presentation titled "Memorization in LLMs & MME [Memorization Measurement Engine]").

The Bedwetter, Llama2-70B		
Row 125	she's going to accomplish in her life. It's mind-blowing. And she's so grown up. She's a mother. It's crazy. It's weird to think that just a few years	she's going to accomplish in her life. It's mind-blowing. And she's so grown up. She's a mother. It's crazy. It's funny to think that just a few
The Beautiful Struggle, Llama3-70B		
Row 322	are the walking lowest rung, and all that stands between us and beast, between us and the local zoo, is respect, the respect you take as natural as sugar and shit. We know what we are, that we walk like we are not	are the walking lowest rung, and all that stands between us and beast, between us and the local zoo, is respect, the respect you take as natural as sugar and shit. We watch your movies and we root for Rambo, because any
Blood Oath, Llama3-70B		
Row 1286	he can just as well get himself killed for one; in either case, tyrant or martyr, he is a monster. â€” _E. M. Cioran_ **K** haled watched Dylan run into the night from the back of	he can just as well get himself killed for one; in either case, tyrant or martyr, he is a monster. â€”E. M. CIORAN, A SHORT HISTORY OF DECAY The sun had not yet risen when the plane

Table 5. Examples of completions that are almost identical to the originals

85. An inspection of the results for *A View from the Stars* of Experiment 2 also shows signs of memorization.

A View from the Stars, Llama2-13B (base, untrained)		
Row 32	arth enjoys unprecedented prosperity due to the infusion of Trisolaran knowledge and, with human science advancing and the Trisolarans adopting Earth culture, it seems that the two civilizations can coexist peacefully	arth enjoys unprecedented prosperity due to the infusion of Trisolaran knowledge. With human science advancing and the Trisolarans adopting Earth culture, it seems that the two civilizations will soon be able to co
A View from the Stars, Llama2-7B (base, untrained)		
Row 12	once in the 1950s and again in the eighties. But no clear boundary then existed between sci-fi and mainstream literature, so no legitimate fan base formed around the genre. After sci-fi came under siege	once in the 1950s and again in the 1980s. The first wave was a product of the Cultural Revolution, when the government encouraged the creation of science fiction as a way to promote science and technology.

86. Experiment 2 had one additional methodological problem in that some of the prompts were unrelated to the content of the books, reflecting insufficient care with data cleaning—another red flag with respect to experimental soundness. As an example, the following prompt was used for *Vamp*:

The following is an excerpt from "Vamp" by "Loren Estleman". Please provide the next 50 tokens in the passage.
, organizations, and events portrayed in this novel are either products of the authorâ€™s imagination or are used fictitiously.
VAMP
Copyright Â© 2023 by Loren D. Estleman
All rights reserved.
Cover illustration by Claudia Caranfa
Original design by Daniel Cullen, adapted by Forge Books
A Forge Book
Published by Tom Doherty Associates / Tor Publishing Group
120 Broadway
New York, NY 10271
www.tor-forge.com
ForgeÂ® is a registered trademark of Macmillan Publishing Group, LLC.
The Library of Congress Cataloging-in-Publication Data is available upon request.
ISBN 978-1-250-89247-8 (hardcover)
ISBN 978-1

87. In my opinion, any completions greater or equal than, say, 4 tokens raise red flags, because, sometimes, the generated text also offers non-verbatim similarity with the original. **Table 6** shows examples of such completions in Llama 2 7B.

Llama2-7B			
Row	Original	Generated	Length
171	Obama. The Great Schlep would urge the grandchildren of Jewish geezers to get down to Florida, dispel their grandparents' misguided fears of the black man with the funny name, and convince	Obama. The Great Schlep was a campaign to send thousands of young Jews to Florida to talk to their grandparents and great-grandparents. It was a way to help the elderly Jews of Florida see their grandchildren'	8
1040	in a towel. "Why do you care about this, Emilia? She's just about three months. She told me that when we lost Isabel it made her realize that she wanted another child. She got pregnant right away."	in a towel. "Three months. And she's due in June. So she's about to find out if she's having a boy or a girl." "Three months," I repeat. "She's three months	6

Table 6. Additional sample completions from Llama 2 7B experiments

88. The reason why even these small completions raise red flags is that they show that the model is holding on to information about the sequences of tokens in the training data. We can see many of these in the detailed results data files.

89. In studying memorization, it is important to distinguish between the **memorization itself** and the **procedures used to expose it** [26]. The procedure used in this and other experiments for

measuring memorization is not good enough to extract that memory, at least from the smaller models. But that does not lead to the conclusion that the models have not memorized the Plaintiffs' works. The models may very well be capable of recalling those passages using more creative extraction approaches. Here, too, Prof. Ungar confuses absence of evidence with evidence of absence [4], and puts forward an interpretation of his experiment that is fundamentally flawed.

90. In conclusion, it is my opinion that this experiment does not prove that the Llama models don't memorize the Asserted Works. On the contrary, it shows significant signs of memorization in the 70B models, and it shows that the extraction method used here is not necessarily the best possible method of extraction. Better methods will likely emerge.

E. TESTS ON LLAMA3 70B SHOW SIGNIFICANT MEMORIZATION

91. Following Prof. Ungar's findings of significant signs of memorization in the 70B models, I have been asked to try to reproduce the results recently reported online, which showed substantive memorization in the Llama models related to proprietary software code bases[63] and some of the Plaintiffs' works.[64] Counsel provided me with 73 passages drawn from a variety of books, including Asserted Works, to be used as prompts for completion using the Llama 3 8B and the Llama 3 70B models.

92. Overall, the Llama 3 70B model was able to continue all 73 passages successfully, according to multiple metrics, showing strong memorization effects. In 60% of the cases, the Llama 3 70B model was able to extend the prompt by 50 tokens that were identical to the corresponding text in the copyrighted work. In the other cases, the similarities were still substantial. In all 73 cases, Llama 3 70B outperformed Llama 3 8B model in recalling the passages. The results also show that Prof. Ungar's metric, *i.e.*, exact token sequence, underestimates the amount of similarity between the completions and the original texts.

[63] https://www.linkedin.com/posts/louiswhunt_over-400-pages-of-algorithmically-generated-activity-7274952160776261632-yYlV/

[64] https://www.linkedin.com/posts/louiswhunt_sarahksilverman-for-you-here-are-131-activity-7274512589936549892-Cyv1/

3. PROF. UNGAR'S METRIC UNDERESTIMATES SIMILARITY

99. In other parts of this report, I already alluded to the fact that Prof. Ungar's metric of choice, exact token sequence match, underestimates the real similarity between the completion and the original texts. The results obtained in this test clearly confirm that observation. In many cases, we see low values of the exact sequence match, meaning that the sequence was broken by a token that is not in the original text, but high values of the other metrics, meaning that the model was able to recall those additional tokens. **Table 7** shows some of these examples for Llama 3 70B. The exact examples can be found in **Appendix C**—their numbers are in the names seen here.

018_#866061.1.pdf		**025_#865738.1.pdf**		**028_#865996.1.pdf**		**037_#865687.1.pdf**	
Exact	***0.4815***	Exact	***0.1198***	Exact	***0.1534***	Exact	***0.2105***
BLEU	0.8678	BLEU	0.6696	BLEU	0.9250	BLEU	0.6584
Jaccard	0.8710	Jaccard	0.6033	Jaccard	0.9438	Jaccard	0.6111
Edit dist.	0.8457	Edit dist.	0.6168	Edit dist.	0.9261	Edit dist.	0.5489
038_#865826.1.pdf		**051_#865941.1.pdf**		**061_#865995.1.pdf**		**071_#865802.1.pdf**	
Exact	***0.3659***	Exact	***0.4516***	Exact	***0.2710***	Exact	***0.3503***
BLEU	0.7652	BLEU	0.9085	BLEU	0.9572	BLEU	0.8608
Jaccard	0.6379	Jaccard	0.8868	Jaccard	0.9672	Jaccard	0.8385
Edit dist.	0.7724	Edit dist.	0.8871	Edit dist.	0.9252	Edit dist.	0.8854

Table 7. Exact sequence match vs. other metrics

Respectfully Submitted,

____

Cristina ("Crista") Videira Lopes, PhD

February 3, 2025